基于多特征的跨语言剽窃检测技术研究 *

刘 刚，胡昱临，李光曦

(哈尔滨工程大学 计算机科学与技术学院, 哈尔滨 150001)

摘 要：针对解决双语剽窃的检测问题，给出了一种跨语言剽窃检测模型。该模型包括了基于多特征选择的跨语言剽窃分类和基于多特征对应的跨语言剽窃检测。该方法主要是根据译者在进行翻译时出现的欧化现象挖掘出常见的译文特征，在对特征进行进一步的特征选择和特征权值的计算后，训练分类器，针对是否存在跨语言剽窃行为进行分类，最后通过 WordNet 进行最后的剽窃确认。通过实验对比和实验分析，分别进行了分类结果和检测结果的验证，证明了所给出的模型的有效性和科学性。

关键词：跨语言剽窃检测；双语特征；欧化现象

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.06.0406

Research on construction technology of cross-language plagiarism detection model based on multi-features

Liu Gang, Hu Yulin, Li Guangxi

(*College of Computer Science & Technology*, *Harbin Engineering University*, *Harbin* 150001, *China*)

Abstract: In order to solve the problem of bilingual plagiarism, this paper constructed a multi-feature-based cross-language plagiarism detection model. This paper firstly analyzes and summarizes the research status of single and double language plagiarism, and proposes a multi-feature-based cross-language plagiarism detection model. The model includes multi-feature-selection-based cross-language plagiarism classification and multi-feature-correspondence–based cross-language plagiarism detection. The results of plagiarism filtering two times is mainly based on the correspondence between translation features and structural features. Finally, the last plagiarism is confirmed by WordNet. In this paper, the transcendental plagiarism model is established, and the results of the classification and the test results are verified by experimental comparison and experimental analysis. The validity and scientificity of the model are proved.

Key words: cross-language plagiarism detection; bilingual feature selection; Europeanized grammar

1 剽窃检测理论

1.1 剽窃的分类

剽窃分为字面剽窃和智能剽窃。其中字面剽窃是比较常见的，它并没有刻意去隐藏所剽窃的内容，只是通过复制粘贴来达到剽窃目的。字面剽窃又分为如下三种：a)精确复制是指不经过任何修改，仅仅对某一段落或者某一整篇文章进行复制；b)相似复制是指通过插入、删减、代替、句子分离或合并等手段进行操作后再复制；c)修改复制是指通过短语重排序或对语法的改变进行修改，然后再据为己所用。总的来说，字面剽窃就是做了很少的改动而没有引用原文章。

而智能剽窃是指用各种方式来试图隐藏和改变原文章。它主要也分为了以下三种方式：a)文本处理是指将文本通过词汇和形态语法进行改变或者通过对概念的归纳、总结和解释的一种剽窃手法；b)翻译是指通过自动翻译（精确翻译、平行语料库等）或者手动翻译将一种语言翻译为另一种语言而没有经过引用，也能够引起剽窃；c)观点剽窃是影响最严重的剽窃，它是指窃取了别人的观点却没有经过引用。

1.2 跨语言文本相似度算法

基于机器翻译是跨语言相似度计算中最直接、最简单的一种方式。它是通过将两种语言统一为同一种形式来进行相似性比较，从而实现跨语言相似度的计算。

基于多语言词典的算法主要是通过双语词典对应来进行匹配的。在 CLIR 和 CLSD 中都有应用，起初由 CLIR 兴起，现发展到 CLSD 并取得了良好的效果。其中比较典型的算法是 CL-CNG（cross-language charater N-Gram）算法[12]。

收稿日期：2018-06-26；**修回日期：**2018-08-16　**基金项目：**黑龙江省博士后科研启动资金资助项目（LBH-Q15031）；黑龙江省教育科学规划课题（GJC1215107）；中央高校基础科研业务费专项资金资助项目（HEUCF180604）

作者简介：刘刚（1976-），男，黑龙江哈尔滨人，副教授，博士，主要研究方向为人工智能、自然语言处理、数据挖掘、机器学习（liugang@hrbeu.edu.cn）；胡昱临（1995-），男，硕士研究生，主要研究方向为人工智能、自然语言处理；李光曦（1990-），男，硕士，主要研究方向为人工智能、数据挖掘.

值得说明的是，CL-CNG 算法只适用于两种相近的语言，但并不适用于汉语和英语这两种区别很大的语言。

这类算法中最为典型的就是跨语言明确语义分析算法（CL-ESA）。它是 ESA 算法的扩展。由 Martin Potthast 等人在 2008 年提出的。

在引入 CL-ESA 算法之前，先介绍 ESA 算法。ESA 算法是单语言之间的语义相似度分析算法，它是由 Wikipedia 作为概念空间，将文本向量用向量空间模型表示，然后使用 TF-IDF 计算其权值，再根据概念空间中概念权值列表表示文本，通过余弦相似度计算两个向量之间的相似性。

设文本 $T=\{w_1,w_2,....,w_x\}$，首先通过 TF-IDF 计算其单词权重 $t=\{v_1,v_2,...,v_x\}$，即表示 w_i 的权重是 v_i，$1\le i\le x$，$\{c_1,c_2,...,c_N\}$ 是概念空间集合，设 w_i 与 c_j 的关联程度是 k_j，那么对应维度 j 的数值可表示为 $\sum_{w_i\in T}v_ik_j$。当计算两段文本相似时，只需将其用 N 维向量表示，然后用余弦定理计算其相似性即可。

同理，CL-ESA 类似[21]，只是将 ESA 算法扩展到跨语言方面，是基于双语 Wikipedia 建立的概念空间，且两者是概念对齐的。其过程如图 1 所示。

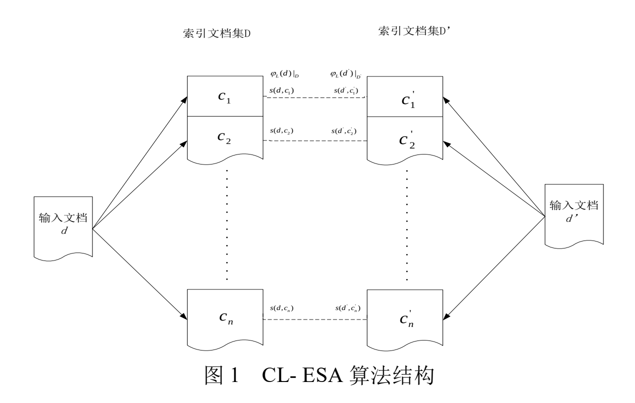


图 1　CL- ESA 算法结构

2　基于多特征选择的跨语言剽窃分类

对于跨语言剽窃来说，首先应确定某篇文章是否存在跨语言剽窃，将存在跨语言剽窃的文章找出，进而才能确定此文章中哪些段落或哪些部分存在跨语言剽窃现象。针对以上问题，本章主要是从具有跨语言剽窃的中文文章中发现并选择其有效的译文特征，给予不同的特征权重，构建具有跨语言剽窃的分类模型，能够对给定的中文文章进行分类，检测其中哪几篇中文文章中可能存在剽窃行为，而哪几篇文章不存在剽窃行为。

2.1　英汉翻译中的欧化现象和翻译体问题发现

翻译体是欧化现象的表现，是指翻译出来的译文有欧化现象或不符合汉语的习惯表达方式，也叫翻译腔、翻译症。文献[20]中将其译为“translationese”。而所谓的欧化，也叫西化，是指语法、文笔、风格或用词受欧洲语文过份影响的中文，影响中尤以英文所造成的最为深刻[48]。欧化中文在语言表达和词语运用上都略显生硬，并且比较容易辨别。

上海外国语大学的李颖玉博士总结了常见的欧化翻译表现形式为以下七种情况：a）外来词及词缀化；b）字母词使用；c）连词增多；d）词类活用；e）助词、数量词、代词滥用；f）长句和冗长句；g）被动句使用增多、标记显化和单一化倾向明显等。

由此可见，在英汉语言相互影响的诸多因素中，词汇和语法的影响比较显著，是区分欧化翻译的最主要的表现形式。中国著名语言学家王力先生曾在文献[7]中的第六章“欧化的语法”一整章都在探讨欧化现象，并对一些“恶意欧化”现象提出了批评。“恶意欧化”现象不仅仅存在于不是以翻译作为本职工作的人，而且对于那些优秀的翻译家而言，也会存在纰漏，何况是对于不同领域的文章。所以，抽象出其中的译文特征来确定某一篇文章存在跨语言剽窃问题是可以解决的，构建并选择合理的译文特征是构建分类模型的关键。

2.2　特征选择——对卡方检验的改进

本文利用卡方检验进行初步的译文特征选择，并且基于 CHI 不足，对 CHI 进行了改进，旨在能够去除一些出现频数较低的且在类别中不稳定的特征，精确找出有效的特征来精确分类。

设类别 c_j 中有 $n_{c_j}(j=1,2)$ 篇文章，特征项 t_i 在每篇文章中出现的频数是 $tf_{i1},tf_{i2},...,tf_{in}$，则特征项 t_i 在 c_j 中的平均频数如式(1)所示。

$$\alpha_{c_j}=\frac{\sum_{i=1}^{n}tf_{ik}}{n_{c_j}} \tag{1}$$

之所以在分母中用所有文章数而不用只存在特征项 t 的文章数，是为了防止低频词只在少部分文章中出现较多，而在绝大多数文章中不出现的情况。这样的情况下将会使频度 α 变大，对稀有词的区分度不高。由此定义特征项 t_i 的频数之差并使之归一化，得

$$\alpha(t_i)=\frac{\alpha_{c_1}-\alpha_{c_2}}{\max(\alpha_{c_1},\alpha_{c_2})} \tag{2}$$

这样的话，将其取值规定到[0,1]区间上。其频度之差越大，越能反映出其区别能力越强，式(2)解决了 CHI 的第一个不足，通过引入 α 来区分特征的频数问题。

针对第二点不足，本文引入了信息熵。信息熵是用来表示随机变量的不确定性的度量，它起源于物理学，用来表征物质状态的参量之一。它主要指任意一种能量在空间中分布的均匀程度。设 X 是一个取有限个值的离散随机变量，其概率分布为

$$p(X=x_i)=p_i,\ i=1,2,...,n \tag{3}$$

则随机变量 X 的熵定义为

$$H(X)=-\sum_{i=1}^{n}p(x_i)\log p(x_i) \tag{4}$$

其中：$0\le H(X)\le\log n$，$H(X)$ 越小，分布越不均匀。

在本文中，需要判断特征 t_i 在指定类别 c_j 中的分布均匀状

况。不妨设 $d_k(0<k\le n)$ 为类别 $c_j(j=1,2)$ 中的第 k 篇文章，则特征项 t_i 在类别 c_j 中信息熵表示如式(5)所示。

$$H(t_i,c_j)=-\sum_{k=1}^{n}\frac{tf(t_i,d_k)}{tf(t_i,c_j)}\log\frac{tf(t_i,d_k)}{tf(t_i,c_j)} \tag{5}$$

其中：$tf(t_i,d_k)$ 表示特征项 t_i 在文章 d_k 中出现的次数；$tf(t_i,c_j)$ 为特征项 t_i 在类别 c_j 中出现的总次数。$H(t_i,c_j)$ 越大，说明分布越均匀，其特征项效果越好。规定如果某特征在该类别中不存在，则 $H(t_i,c_j)=1$。

定义

$$H(t_i)=\frac{H(t_i,c_1)}{H(t_i,c_2)} \tag{6}$$

其中：$H(t_i,c_j)$ 为特征项 t_i 在类别 c_j 中的信息熵。当在 c_1 类中越稳定，在 c_2 类中越不稳定，则 $H(t_i)$ 的值越大，越能代表剽窃类 c_1。这样，对于所有的特征项 $t_1,t_2,\ldots,t_n$ 对应的 $H(t_i)$，将其进行归一化，得

$$Ho(t_i)=\frac{H(t_i)}{\max(H(t_1),H(t_2),\ldots,H(t_n))} \tag{7}$$

显然，$Ho(t_i)\in[0,1]$。

综上所述，定义新的 CHI 方式：

$$CHI_{new}(t_i,c)=k_1P_{t_i}+k_2\alpha(t_i)+k_3Ho(t_i) \tag{8}$$

其中：P_{t_i} 为 $\chi^2(t_i,c)$ 值查询卡方分布的临界值表得到的概率；$\alpha(t_i)$ 为特征项 t_i 平均频数之差；$Ho(t_i)$ 为特征项 t_i 在类别中的信息熵。后两者都进行了归一化处理，故三者都在[0,1]内。k_1,k_2,k_3 为每个因素的权重。$CHI_{new}(t_i,c)$ 的值越大，说明该特征区分度越高，是有效的，其值越小说明该特征区分度越低，是无效的。

2.3　SVM 模型训练

本文采用非线性支持向量机作为模型，选用 RBF（radial basis function，径向基函数）作为核函数，最后经过学习要得到分类决策函数是

$$f(x)=\mathrm{sign}(\sum_{i=1}^{N}\alpha_i^* y_i K(x,x_i)+b^*) \tag{9}$$

其中：RBF 为

$$K(x_i,x_j)=\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2}) \tag{10}$$

其具体的 SVM 分类模型构建及求解方法如算法 1 所示。

算法 1　基于译文特征的 SVM 模型构建与求解算法

输入：训练数据集 D 以及特征 T。

输出：判断是否剽窃的分类模型 $f(x)$。

$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$，

$x_i\in\chi=R^M,\ y_i\in Y=\{+1,-1\},\ i=1,2\ldots N$

选取参数 C，用 RBF 代替内积，得到 SVM 的对偶问题

取初值 $\alpha^{(0)}=0$，令 $k=0$；

while(当α存在不满足KTT条件的变量)

选取优化变量 $\alpha_1^{(k)},\alpha_2^{(k)}$；

将对偶问题转换为式(2-21)的形式；

得到最优解 $\alpha_1^{(k+1)},\alpha_2^{(k+1)}$，并更新 α 为 $\alpha^{(k+1)}$；

　if(α在精度ε内满足KTT条件)

　　break;

　end if　//如果满足 KTT 条件就跳出循环

$k++$

end while

取 $\hat{\alpha}=\alpha^{(k+1)}$

这样，根据求出的最优解 α^* 来计算 b^*

由 α^* b^* 可得其分类模型 $f(x)$

算法 1 说明了基于多种译文特征的 SVM 分类器的构建和求解过程。首先将原始问题转换为对偶问题，然后运用第二章提到的 SMO 算法对不满足约束条件的变量进行更新，直到所有变量都满足 KTT 条件，进而根据求解出来的最优解来求得分类模型。

3　基于多特征对应的跨语言剽窃检测

进行分类模型构建后，给出一篇中文文章，可以判断是否进行了跨语言剽窃。在确认该篇文章是存在跨语言剽窃时，需要进一步确认具体剽窃了哪一篇文章，进而精确到剽窃了哪一个段落。本章基于上述问题，提出了基于多特征对应的跨语言剽窃检测方法，本章是上一章的延续，通过上一章得到的剽窃候选集，将进一步精确分析其特征对应情况，确认出剽窃的具体位置。

3.1　基于译文特征对应的剽窃结果一次过滤

根据英汉翻译中的欧化现象和翻译体问题，构建出了在中文文章中存在的译文特征，并且根据译文特征找出了可能存在剽窃的中文文章。换一种思路来想，中文中的译文特征如果对应到英文中也是存在的，可以根据中英文译文特征出现的位置来进一步确定具体的剽窃结果。

每个特征在每句话中的特征表示进行加权，得到两个 n 维有序向量，这两个向量即为 n 个特征在要比较的中英文段落中的特征表示，计算出这两个向量的欧氏距离即为段落之间的距离。距离越短，说明这两个段落越相似。

例 1　图 2 和 3 分别是中文段落及其对应的英文段落。

[近年来，在许多行业中存在对尖端的基于云的应用的巨大需求。][我们在文件中提出了共享磁盘云数据库架构作为基础，可以开发智能数据存储管理系统以丰富基于云的 Web 应用程序。][这种提出的架构的重要特征是单拷贝数据一致性，动态负载平衡和高基准性能。][基于软件层，已经指出了用于推广 SaaS 概念的智能数据管理系统，其提出了用于普及云环境的成本有效的解决方案。]

图 2　中文段落示例

[In recent years, there is tremendous demand of cutting-edge cloud-based applications in many of the industries.][We have proposed in the paper a shared disk cloud database architecture as the basis on which an intelligent data storage management system can be developed for enriching cloud-based web applications.][Important features of this proposed architecture are single copied data consistency, dynamic load balancing and high benchmark performance.][Based on the software layer, an intelligent data management system for popularizing the concept of SaaS has been pointed out suggesting a cost-effective solution for popularizing the cloud environment.]

图 3　英文段落示例

首先根据选择出来的特征进行中英文特征的对应。在中文段落中，特征对应的矩阵为

$$\begin{bmatrix} 0 & \ldots & 0 & 0 & 0 & \ldots & t_{13} & \ldots t_{16} \ldots 0 & 0 & \ldots & 0 \\ 0 & \ldots & t_6 & 0 & 0 & \ldots & t_{13} & \ldots 0 \ldots 0 & t_{24} & \ldots & 0 \\ 0 & \ldots & 0 & 0 & t_8 & \ldots & 0 & \ldots 0 \ldots 0 & t_{24} & \ldots & 0 \\ 2t_1 & \ldots & 0 & 0 & 0 & \ldots & 0 & \ldots 0 \ldots 0 & 0 & \ldots & 0 \end{bmatrix}$$

在英文段落中，特征对应的矩阵为：

$$\begin{bmatrix} 0 & \ldots & 0 & 0 & 0 & \ldots & 2t_{13} & \ldots t_{16} \ldots 0 & 0 & \ldots & 0 \\ 0 & \ldots & t_6 & 0 & 0 & \ldots & t_{13} & \ldots 0 \ldots 0 & t_{24} & \ldots & 0 \\ 0 & \ldots & 0 & 0 & t_8 & \ldots & 0 & \ldots 0 \ldots 0 & t_{24} & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & \ldots 0 \ldots 0 & 0 & \ldots & 0 \end{bmatrix}$$

由于篇幅所限，英文中出现的其他特征值在此没有表示，但在实际计算中不能忽略。在这里，t_i 是第三章计算出来的每个特征的权重值，是常数。与此同时，在将矩阵转换成段落向量之前，需要确定每一句的权值，将段落表示成矩阵即为

$$\{0.62t_1,\ldots,0.19t_6,0,0.19t_8,\ldots,0.5t_{14},\ldots,0.19t_{17},\ldots,0.19t_{25},0.38t_{26},\ldots,0\}$$

$$\{0,\ldots,0.19t_6,0,0.19t_8,\ldots,0.5t_{14},\ldots,0.19t_{17},\ldots,0.19t_{25},0.38t_{26},\ldots,0\}$$

根据公式计算两者之间的欧氏距离 d 即可。

本节利用特征进行了中英文特征的对应，过滤了不符合特征对应的段落，进而将剽窃的结果的范围大大缩小。

3.2　基于结构特征对应的剽窃结果二次过滤

算法 2 基于结构特征的段落过滤算法

输入：　中文剽窃段落 P，初步剽窃结果 E。

输出：　筛选后的剽窃结果。

给定阈值 $l,l_n,l_v,l_{adj},l_{adv}$，给定第 i 篇中文剽窃段落

对于每一篇保留段落 E_j

$if\ \ P_l-E_{jl}>l \,||\, P_{l_n}-E_{jl_n}>l_n \,||\, P_{l_v}-E_{jl_v}>l_v$

$||\, P_{l_{adj}}-E_{jl_{adj}}>l_{adj} \,||\, P_{l_{adv}}-E_{jl_{adv}}>l_{adv}$

该段落不符合条件；

Else

$list.add(j);$ //保存该剽窃段落

$map.put(i,list);$ //将所有通过筛选的剽窃结果放入以 P 为 key 的 map 中；

返回 map 值

本文选取了句子的长度、句子中名词的长度、句子中动词的长度、句子中形容词的长度、句子中副词的长度五种结构特征，用来对剽窃候选集进行进一步筛选和过滤，如算法 2 所示。

算法 2 给定了五种特征的阈值，给定一篇中文剽窃段落和上一小节筛选出来的一次过滤的剽窃结果一一进行比较，如果某个特征超出特定阈值，则将其从其剽窃结果中进行过滤，过滤之后剩余的段落即为二次过滤后的剽窃结果。

3.3　基于 wordnet 的剽窃结果最终认定

在进行两次过滤之后，得到了最终剽窃结果。剽窃结果中可能只有一个段落，即已经找出中文段落所剽窃的英文段落，只是从语义上待进一步确认；也可能是多个段落，需要从多个段落中精确找出剽窃的段落。鉴于此，本节引入基于 WordNet 的跨语言文本相似度的计算方法[22]来进行最终结果的确认。

在进行名词消歧后，每一个名词都能得到一个有用的指纹序列，但并不是所有的名词都是有用的。有些名词出现频率很低，不具有典型性，诸如此类的都需要进行过滤，留下分辨率较大的指纹来进行相似度的计算。

本文采取与 *TF-IDF* 计算权重类似的方法来选取指纹。对于一些多次出现的名词，即它的 *TF* 大，给予保留，而对于逆文档频率 *IDF* 的选取需要依赖于数据集。因此本文基于 WordNet 的同义词数据集在树型结构中的深度作为过滤特征集的条件[11]。深度越浅，该节点代表的含义越弱。因此把低 100 全为 0 的指纹其进行了过滤，剩余的指纹即为选取出来的进行相似度计算的指纹。

在进行名词语义哈希、名词消歧、指纹选取后，得到了正式的哈希特征序列。设语言 L 的输入文本 d 和语言 L' 的输入文本 d' 的特征序列分别为

$$F(d)=\{\varphi(s_1),\varphi(s_2),\ldots\}$$

$$F(d')=\{\varphi(s_1'),\varphi(s_2'),\ldots\}$$

则通过 Dice 系数来计算文本 d 和文本 d' 的相似度，如式(11)所示。

$$sim(d,d')=\frac{2\times|F(d)\cap F(d')|}{|F(d)|+|F(d')|} \tag{11}$$

这样，便可得到两者的相似度。

4　实验及验证

本实验的相关环境如下：

实验平台：Windows 7(64 位)；

处理器：Pentium (R) Dual-Core @2.50 GHz；

内存：4.00 GB；

实验环境：MyEclipse, WinPython-64bit-2.7.10.3；

开发语言：Java，Python；

实验数据：本实验的实验数据分为训练数据集和测试数据集两部分。

a）训练数据集的数据的正样本来自 Springer 里面的 Computer Science 学科下的 Chapter 下的 3 500 篇文章，通过自动翻译为中文文本作为训练集中的正样本。训练集数据的负样本来自于从中国知网的计算机软件与计算机应用类别中的中国学术期刊网络出版总库，里面包含着由《计算机学报》、《软件学报》等国内著名学报的期刊，选取 2 800 篇中文文章作为训练集负样本。

b）测试数据集为 Springer 中的 100 英文文章和它们的中文翻译以及 50 篇知网的中文文章。

4.1　第一次过滤

在对文本进行预处理后，如前述方法将文本中的特征提取出来，将一些出现频度偏低的特征去掉后，将符合的特征进行信息熵的计算，得出每个特征项的的稳定程度；接下来，需要确定三个权重参数 k_1,k_2,k_3 的值。根据结果确定 3 个参数的值复杂度太大，显然是不可取的。通过人工排序和算法 1 选择最优参数为 $k_1=0.04$， $k_2=0.78$， $k_3=0.13$。其对比结果如图 4 所示。

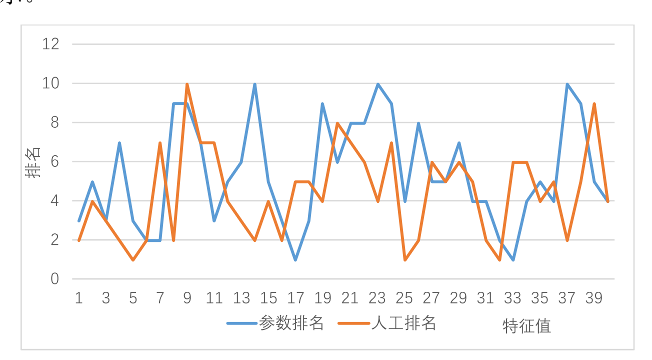


图 4　选定参数对比分析图

在得到特征权重之后，将训练集中的文章进行特征表示，然后运用 SVM 进行分类器训练，得到分类模型。

基于译文特征做剽窃分类的文章很少，本文将特征选择及特征赋予的权重后作训练得出的三个评价指标，与特征选择后及特征赋予的权重之前作训练得出的三个评价指标和文献[18]提供的特征训练所得到的三个指标作对比，用本文的训练数据集和测试数据集分别进行封闭测试和开放测试，其结果对比如图 5 和 6 所示。

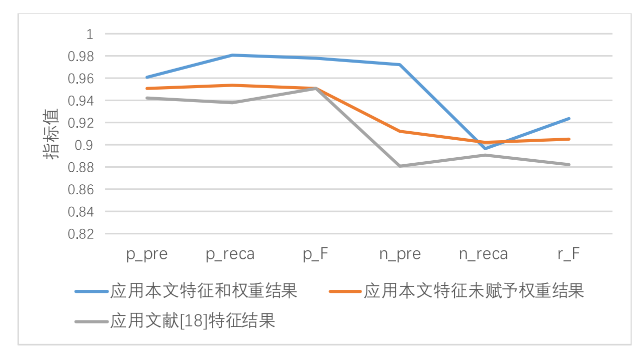


图 5　封闭测试评价指标对比

从图中可以看出，在封闭测试中，本文方法与文献[18]相比，除了在非剽窃文本的召回率上与文献[18]的方法持平以外，在其余指标上有了很大提高，综合对比 F 值也有优势。而在开放测试中，该优势更加明显，各个指标均领先于其他指标。所以，本文针对跨语言剽窃中特征的选取准确性有了很大提高，证明了本文特征选取方法的有效性所在。

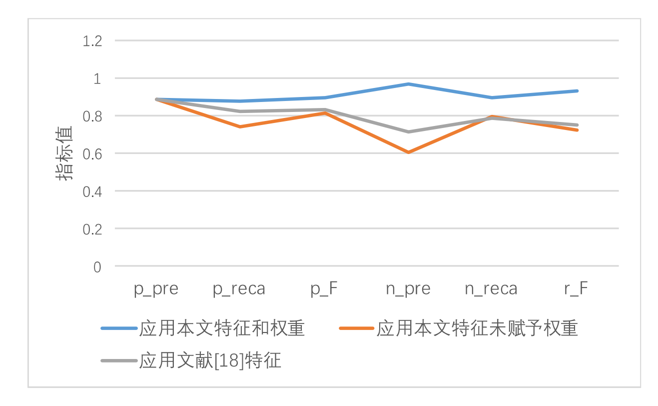


图 6　开放测试评价指标对比

4.2　第二次过滤

对于一个中文段落的多个英文段落，将不符合条件的全都过滤掉。在对比 1 000 个段落后，有 749 个段落经过两次过滤后只保留了一个可疑段落，其中有 736 个段落精确匹配到其剽窃的段落，仅 13 个段落出现了匹配错误的情况。在剩余 251 个段落候选集中，有 24 个段落经过两次过滤没有可疑段落与之匹配，有 227 个段落有多个可疑段落与之匹配。图 7 展示了上述结果。由图可见此时正确率已达 74%。而在 227 篇与多个结果匹配的段落中，需要筛选出与具体的剽窃段落，这将借助 WordNet 词典完成最终结果的确认工作。

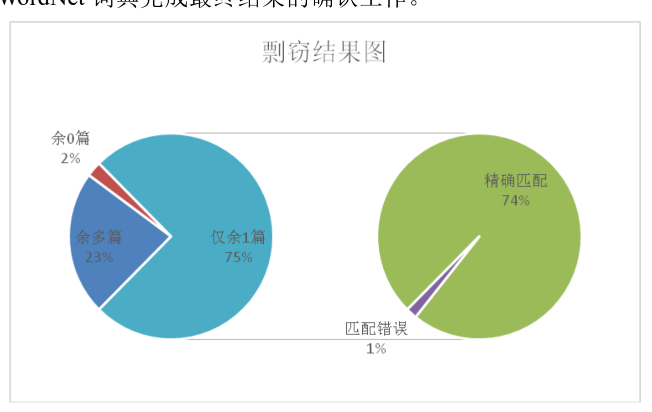


图 7　剽窃结果

经统计，在验证的 227 个段落中，有 220 个段落实现了剽窃结果的准确对应，仅有 7 个段落的筛选错误，归其原因，在 WordNet 计算相似度时出现误差，正确段落并没有得到最大的相似度。但所剽窃段落在过滤后存在其可疑段落中，从侧面说明了结果有效性。

在数据集上作本文基于特征对应的剽窃结果两次过滤，然后用基于 WordNet 的跨语言相似度检测，与文献[18]直接基于 WordNet 的跨语言相似度检测，其准确率、召回率、F 值对比如图 8 所示。

由图中可以看出，经过本文实验，精确率和召回率都进行了提升。归其原因，两次过滤将一些在词义上相似但译文和结

构特征差别大的段落都进行了过滤，只留下了一些译文和结构特征差别小但词义也不是很相近的段落，所以精确度有了很大提高。这也验证了本文理论的有效性。

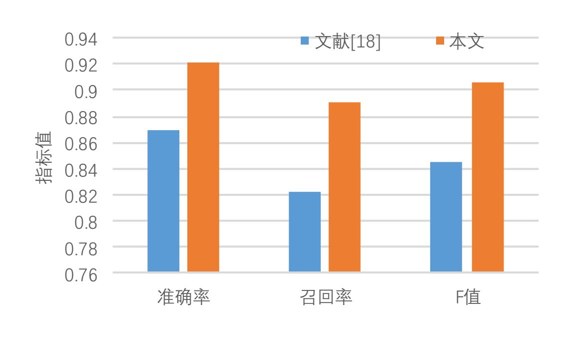


图 8　结果分析图

5　结束语

本文提出的方法跨越了语言与语法之间的不一致问题，从一个新的角度进行了剽窃检测。但正如前面所说的，跨语言剽窃检测才刚刚起步，还存在着诸多的不足，需要不断去完善修改。首先，在语料库选取上，语料库的质量将直接影响最后的分类训练结果，所以未来需要在建设高质量的语料库上下足功夫；其次，在特征构建时，需要进一步完善和挖掘特征，实现对翻译特征的自动挖掘也是未来的研究重点之一；最后，在效率上，需要更加注重效率问题，尤其在面对大数据集训练时，这也是未来需要重点研究的内容。

参考文献：

[1] 张阁阁，孙梅影. 浅谈科学共同体内部的伦理问题——以高等院校和科研机构为例 [J]. 卷宗, 2015 (6): 622-625. (Zhang Gege, Sun Meiying. Ethical issues within the scientific community: taking institutions of higher learning and scientific research institutions as examples [J]. JuanZong, 2015 (6): 622-625.)

[2] Brassil J T, Low S, Maxemchuk N F, *et al.* Electronic marking and identification techniques to discourage document copying [J]. IEEE Journal on Selected Areas in Communications, 1995, 13 (8): 1495-1504.

[3] 康存辉. 道德治理视阈下的学术不端检测是与非 [J]. 武汉纺织大学学报, 2015 (2): 74-76. (Kang Cunhui. Detection of academic misconduct from the perspective of Moral Governance [J]. Journal of Wuhan Textile University, 2015 (2): 74-76.)

[4] 邹杜，陈育青，张凌. 基于语义匹配的抄袭检测方法 [J]. 华南理工大学学报：自然科学版, 2013, 41 (7): 131-136. (Zou Du, Chen Yuqing, Zhang ling. Method of plagiarism detection based on semantic matching [J]. Journal of South China University of Technology: Natural Science Edition, 2013, 41 (7): 131-136.)

[5] 张伟. 基于 n-gram 的中文文本复制检测研究 [D]. 长沙：湖南大学, 2014. (Zhang Wei. Research on Chinese text copy detection based on n-gram [D]. Changsha: Hunan University, 2014.)

[6] 夏志明，刘新. 一种基于语义的中文文本相似度算法 [J]. 计算机与现代化, 2015 (4): 6-9. (Xia Zhiming, Liu Xin. A Chinese text similarity algorithm based on semantics [J]. Computer and Modernization, 2015 (4): 6-9.)

[7] 张贤坤，张倩. 基于本体的综合加权案例相似度算法研究 [J]. 算法研究探讨, 2017 (2): 422-425. (Zhang Xiankun, Zhang Qian. Research on ontology based comprehensive weighted case similarity algorithm [J]. Application Research of Computers, 2017 (2): 422-425.)

[8] 谢松山，唐雁. 基于左归词频向量空间模型的中文文本抄袭检测算法 [J]. 西南大学学报：自然科学版, 2015, 37 (5): 158-161. (Xie Songshan, Tang Yan. Chinese text plagiarism detection algorithm based on left left word frequency vector space model [J]. Journal of Southwest University: Natural Science Edition, 2015, 37 (5): 158-161.)

[9] 朱群燕. 基于可比语料库的跨语言信息检索研究 [D]. 武汉：华中师范大学, 2015. (Zhu Qunyan. Research on cross language information retrieval based on comparable corpus [D]. Wuhan: Huazhong Normal University, 2015.)

[10] Franco-Salvador M, Gupta P, Rosso P. Knowledge graphs as context models: improving the detection of cross-language plagiarism with Paraphrasing [C]// Proc of Bridging Between Information Retrieval and Databases. Berlin: Springer, 2014: 227-236.

[11] Franco-Salvador M, Rosso P, Montes-Y-Gómez, *et al.* A systematic study of knowledge graph analysis for cross-language plagiarism detection [J]. Information Processing & Management, 2016, 52 (4): 550–570.

[12] 彭哲. 跨语言文本相关性检测技术研究 [D]. 长沙：中南大学, 2014. (Peng Zhe. Research on cross language text correlation detection technology [D]. Changsha: Central South University, 2014.)

[13] 刘娇，崔荣一，赵亚慧，等. 跨语言文献相似度的分析方法 [J]. 延边大学学报：自然科学版, 2016, 42 (2): 151-155. (Liu Jiao, Cui Rongyi, Zhao Yahui, *et al.* An analysis method of cross-lingual literature similarity [J]. Journal of Yanbian University: Natural Science, 2016, 42 (2): 151-155.)

[14] 张靖. 面向高维小样本数据的分类特征选择算法研究 [D]. 合肥：合肥工业大学, 2014. (Zhang Jing. Classification feature selection algorithm for high-dimensional small sample data [D]. Hefei: Hefei Polytechnic University, 2014.)

[15] Mcnamee P, Mayfield J. Character n-gram tokenization for European language text retrieval [J]. Information Retrieval, 2014, 7 (1-2): 73-97.

[16] 蒲晓燕.“英式汉语”称谓、英文译名及定义辨析 [J]. 南昌教育学院学报, 2015 (12): 163-180. (Pu Xiaoyan. An analysis of the titles, English translations and definitions of"English Chinese" [J]. Journal of Nanchang College of Education, 2015 (12): 163-180.)

[17] Nitto E D, Matthews P, Petcu D, *et al.* Model-driven development and operation of multi-cloud applications [M]. Berlin: Springer International Publishing, 2017.

[18] 杨倩茹. 基于指纹融合的跨语言剽窃检测技术研究 [D]. 哈尔滨：哈尔滨工程大学, 2016. (Yang Qianru. Research on cross language plagiarism detection technology based on fingerprint fusion [D]. Harbin: Harbin

Engineering University, 2016.)

[19] Franco-Salvador M, Gupta P, Rosso P. Cross-language plagiarism detection using a multilingual semantic network [C]// Proc of European Conference on Advances in Information Retrieval. Berlin: Springer, 2013: 710-713.

[20] 罗远胜，王明文，勒中坚，等. 跨语言信息检索中的双语主题相关模型 [J]. 小型微型计算机系统, 2013, 34 (12): 2758-2763. (Luo Yuansheng, Wang Mingwen, Le Zhongjian, *et al.* Bilingual topic correlation model in cross language information retrieval [J]. Journal of Chinese Computer Systems, 2013, 34 (12): 2758-2763.)

[21] Narducci F, Palmonari M, Semeraro G. Cross-lingual link discovery with TR-ESA [J]. Information Sciences, 2017, 394-395: 68-87.

[22] Gamallo P, Pereira-Fariña M. Compositional semantics using feature-based models from wordNet [C]// Proc of Workshop on Sense. 2017.